Vol.2010-FI-98 No.8
Vol.2010-DD-75 No.8
2010/3/5

距離尺度学習を用いた多次元尺度構成法 (MDS) による可視化結果のインタラクティブな操作方法の提案

吉岡　真治[†1]　伊藤　正彦[†2]
Michéle Sebag [†3]　Jean-Daniel Fekete [†4]

多次元尺度構成法 (MDS) は、高次元のデータを 2 次元で表現する手法である。しかし、ユーザが初期の出力結果に満足できない場合が存在する。この問題に対し、本発表では、ユーザにとって分類に役立つ次元を学習する距離尺度学習の手法を応用することによって、MDS の結果をインタラクティブに操作する方法を提案する。また、このシステムを論文の類似度を元にして作成した複数研究者の関係を可視化するシステムに適用した結果について紹介する。

Interactive Operation of MDS Visualization Results with Distance Metric Learning

Masaharu Yoshioka [†1], Masahiko Itoh [†2]
Michéle Sebag [†3] and Jean-Daniel Fekete [†4]

Multi Dimensional Scaling (MDS) is one of the popular method for visualizing multidimensional data into 2D world. However, the user may not satisfy the visualization results. In this paper, we propose a novel interactive method for modifying MDS visualization results by using the concept of distance metric learning. We also demonstrate our researcher visualization system based on this algorithm.

1. はじめに

多次元尺度構成法 (Multi Dimensional Scaling：MDS) は、文書データや画像データなどの様々なデータをデータ間の類似性などに基づいて表示する方法として、良く用いられる方法である[1)]。この多次元尺度構成法では、高次元空間から 2 次元空間に写像をした際に、高次元空間における各データ間の距離と 2 次元空間における各データ間の距離の差を最小にするといった、データ群が持つ数学的な性質に注目した写像が行われる。

しかし、このような数学的な性質に基づいた写像が、一般ユーザの持つ類似度の直感と異なる場合が考えられる。この様な場合に、MDS の結果をインタラクティブに修正する方法などが提案されている[1)–3)]。これらのインタラクティブ操作の多くは、元の高次元における距離を所与のものとした操作を提案している。

一方、機械学習の分野では、高次元の属性空間において、分類に役立つ次元とそうでない次元を判断することによって、有効な距離尺度を計算するという距離尺度学習 (Distance Metric Learning) という考え方が提案されている[4)]。

本研究では、この距離尺度学習 (Distance Metric Learning) の考え方を取り入れることによって、MDS のインタラクティブな操作を実現する方法を提案する。また、本手法の有効性を検証するために、著者の類似度を論文のアブストラクトの類似度に基づいて可視化する、複数研究者の関係を可視化するシステムに適用した結果について紹介する。

2. k 近傍分類法のための距離尺度学習

k 近傍分類法 (k-nearest neighbors (kNN) method) は、分類問題を解くための良く知られたアルゴリズムである。このアルゴリズムでは、k 個の最近傍のオブジェクトにおいて、最も一般的なラベルをそのオブジェクトに割り当てることによって分類問題を解く。この手法では、データ間の距離が非常に重要な値となるため、適切な距離尺度が与えられた場合に

†1 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology,Hokkaido University
†2 東京大学生産技術研究所
Institute of Industrial Science, The University of Tokyo
†3 パリ南大学 CNRS
CNRS, Université Paris Sud
†4 パリ南大学 INRIA
INRIA, Université Paris Sud

Vol.2010-FI-98 No.8
Vol.2010-DD-75 No.8
2010/3/5

は、分類がうまくいくが、不適切な距離尺度を与えた場合には、分類がうまくいかないといったことが起こる。

Weinberger ら[4] は、この k 近傍分類法の分類性能を向上させるための距離尺度の学習アルゴリズムを提案している。このアルゴリズムでは、以下の 2 つの基準を可能な限り満足させる距離尺度を、有用な尺度として学習する。

Push k 個の最近傍のデータが、出来る限り同じラベルに属する。

Pull 異なるクラスに属するデータが近くに存在する場合には、同じラベルのデータよりも与えられたマージン以上の距離だけ遠ざける。

つまり、Push の基準により、同じラベルに属するオブジェクトを近くに配置し、Pull の基準により、異なるラベルに属するオブジェクトを遠くに配置することのよって、k 近傍分類法の分類誤りが少ない距離尺度を得ようという考え方である。

このアルゴリズムでは、初期状態として与えられた多次元空間に対して、上記の基準を満たす線型変換行列 L を求める問題として定式化する。この L を用いて、二つのオブジェクトを表す $\vec{x_i}, \vec{x_j}$ ベクトル間の距離は以下の式を用いて計算する。

$$D(\vec{x_i}, \vec{x_j}) = ||L(\vec{x_i} - \vec{x_j})|| \tag{1}$$

この目的を実現するために、push と pull の基準に対応する次のようなコスト関数を設定し、コスト関数の最小化することにより、L を求める。

$$\epsilon_{pull}(L) = \sum_{j \rightsquigarrow i} ||L(\vec{x_i} - \vec{x_j})||^2 \tag{2}$$

$$\epsilon_{push}(L) = \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il}) \left[1 + ||L(\vec{x_i} - \vec{x_j})||^2 - ||L(\vec{x_i} - \vec{x_l})||^2\right]_+ \tag{3}$$

$$\epsilon(L) = (1 - \mu)\epsilon_{pull}(L) + \mu\epsilon_{push}(L) \tag{4}$$

ただし、

- y_{il} は x_i と x_j が同じラベルに属するときのみ $y_{il} = 1$ とし、そうでない場合は、$y_{il} = 0$ とする。
- $[z]_+ = max(z, 0)$ は、一般的なヒンジロス関数である。
- $j \rightsquigarrow i$ は $\vec{x_j}$ が $\vec{x_i}$ にとっての k 最近傍であることを示す。

このコスト関数を最小化するためには、k 最近傍の同じラベルのオブジェクト間の距離が小さくなること (Push) と、k 最近傍に含まれてしまう異なるラベルのオブジェクトを遠ざけること (Pull) をバランス良く行うことが求められる。

この最適化問題を効率よく解くために、線型変換行列 L に対応するマハラノビス距離を表す行列である M を導入する。

$$D_M(\vec{x_i}, \vec{x_j}) = (\vec{x_i}.\vec{x_j})^T M (\vec{x_i}.\vec{x_j}) \tag{5}$$

この行列 M を用いることにより、この最適化問題が以下の半正定値問題として扱うことができ、半正定値問題プログラム (semidefinite programming problem：SDP) を用いて最適値を求めることができるようになる。

$$\epsilon(M) = (1 - \mu) \sum_{j \rightsquigarrow i} D_M(\vec{x_i}, \vec{x_j}) + \mu \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il}) [1 + D_M(\vec{x_i}, \vec{x_j}) - D_M(\vec{x_i}, \vec{x_l})]_+ \tag{6}$$

ただし、

$$(\vec{x_i}.\vec{x_l})^T M (\vec{x_i}.\vec{x_l}) - (\vec{x_i}.\vec{x_j})^T M (\vec{x_i}.\vec{x_j}) \leq 1 - \xi_{ijl} \tag{7}$$

$$\xi_{ijl} \leq 0 \tag{8}$$

$$M \succeq 0 \tag{9}$$

この条件の最後にある $M \succeq 0$ は M が半正定値であるための条件である。この定式化を用いることによって、M を求める問題を SDP のパッケージを用いて解くことが出来るようになる。ただし、一般の SDP のパッケージをそのまま利用するのでは、速度が十分でないため、劣勾配法 (subgradient) を用いて計算を行う。

表記を簡略化するために、$C_{ij} = (\vec{x_i} - \vec{x_j})(\vec{x_i} - \vec{x_j})^T$ という記法を用いて、式 6 を以下のように書き換える。

$$\epsilon(M_t) = (1 - \mu) \sum_{j \rightsquigarrow i} tr(M_t C_{ij}) + \mu \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{il}) [1 + tr(M_t C_{ij}) - tr(M_t C_{il}))]_+ \tag{10}$$

この結果、式 10 は M_t に関して線型である。また、i 番目のオブジェクトの k 最近傍である j 番目のオブジェクトに対し、別のラベルを持つオブジェクトでマージンよりも内側に存在するオブジェクトの番号 l の三つ組 (pull の条件を満たすデータの組合わせ) から構成される集合 $(i, j, l) \in N^t$ を用いて、今回の問題 $\epsilon(M_t)$ における勾配 G_t は以下の式で計算できる。

Vol.2010-FI-98 No.8
Vol.2010-DD-75 No.8
2010/3/5

$$G_t = \frac{\partial \epsilon(M_t)}{M_t} = (1-\mu)\sum_{j \rightsquigarrow i} C_{ij} + \mu \sum_{i,j \rightsquigarrow i} \sum_{l} (1-y_{il})(C_{ij} - C_{il}) \quad (11)$$

この勾配を逐次的に計算していくことにより、ϵ を最小化する M を計算する。

Weinberger らは、本手法を幾つかの分類データに適用することにより、k 最近傍分類法の分類性能が向上することを確認している。

3. 多次元尺度構成法による可視化結果へのインタラクティブな操作

多次元尺度構成法 (MDS) は、高次元の情報として表されるデータを 2 次元空間に写像して表示する方法である。この写像の際に用いられる基準としては、各々のデータ間の高次元空間における距離の情報を出来る限り保存するといった数学的な基準が用いられる。

このような基準は、データ全体の散らばり具合いの様子を出来る限り保存したいといった観点からは有用であるが、一般的なユーザにとって、その写像結果は必ずしも満足のいくものではないということが起こりうる。

このような問題に対し、MDS の結果をユーザの位とに基づいて修正するという幾つかの方法が提案されている[1)–3)]。これらのインタラクティブ操作の多くは、元の高次元における距離を所与のものとした操作を提案している。

これに対し、本研究では、前節で紹介した距離尺度学習の枠組みを導入したインタラクティブなオペレーションを提案する。この枠組みでは、ユーザが近くに配置したいと考えるオブジェクトに対してラベルを与えることによって、新しい距離尺度を学習し、MDS の結果を更新する。以下にこのインタラクティブな操作を実現するための大まかな手順について説明する。

(1) 初期のオブジェクトの情報から、各々のオブジェクト間の距離を計算し、MDS の初期配置を決定する。
(2) ユーザは、類似しているオブジェクトと類似していないオブジェクトを選択し、システムに伝える。
(3) システムは、類似しているオブジェクトには同じラベルを、類似していないオブジェクトには異なるラベルを設定し、距離尺度学習を行う。
(4) 新しいオブジェクト間の距離を学習の結果得られたマハラノビス行列を用いて計算し、MDS の結果を更新する。結果に満足しない場合には、手順の 2 番目に戻る。

3.1 多次元尺度構成法のための距離尺度学習

2 節で述べた距離尺度学習では、学習を行うにあたり多くのデータに対してラベルがついていることが前提となったアルゴリズムであった。これに対し、本研究で考えるようなインタラクティブな操作を前提とする場合には、このように多くのデータにラベルがついていることは想定しにくい。よって、ラベルがついているデータ数が少ないことを前提とした学習のアルゴリズムを設定する必要がある。

具体的には、既存手法では、k 最近傍のデータのみに注目していたが、本研究では、同じラベルのついている全てのデータに注目して学習を行うことにした。この考え方に基づき、既存手法の push,pull を以下のように修正した。

Push 全ての同じラベルに属するデータが、出来る限り近くに集まる。

Pull 異なるクラスに属するデータが同じクラスのデータよりマージンを考慮した上で近くに存在する場合には、同じラベルのデータよりも与えられたマージン以上の距離だけ遠ざける。

このとき、ラベルのついていないデータについては、push や pull の対象にならない事とした。

この考え方に基づいて、既存手法のコスト関数を以下のように修正した。

$$\epsilon'_{pull}(M_t) = \sum_i \sum_{j \neq i} s_{ij} D_{M_t}(\vec{x_i}, \vec{x_j}) \quad (12)$$

$$\epsilon'_{push}(M_t) = \sum_i \sum_{j \neq i} \sum_l s_{ij}(1 - sr_{il}) [1 + D_{M_t}(\vec{x_i}, \vec{x_l}) - D_{M_t}(\vec{x_i}, \vec{x_l})]_+ \quad (13)$$

$$\epsilon'(M_t) = (1-\mu)\epsilon'_{push}(M_t) + \mu\epsilon'_{pull}(M_t) \quad (14)$$

ただし、

- s_{ij} は、x_i と x_j が同じラベルを共有している場合にのみ $s_{ij} = 1$ となり、それ以外の場合には 0 となる。
- sr_{ij} は、x_i と x_j が異なるラベルを共有していない場合にのみ $s_{ij} = 1$ となり、それ以外の場合には 0 となる。

このコスト関数 $\epsilon'(M_t)$ に対する勾配 G'_t は以下の式で計算される。

Vol.2010-FI-98 No.8
Vol.2010-DD-75 No.8
2010/3/5

$$G'_t = \frac{\partial \epsilon'(M_t)}{M_t} = (1-\mu)\sum_i \sum_{j \neq i} C_{ij} + \mu \sum_i \sum_{j \neq i} \sum_l (C_{ij} - C_{il}) \quad (15)$$

この勾配を用いることにより、2 節と同様に、距離尺度学習を行う。

ただし、2 節のように、繰り返し計算を行って最適解を計算する方法では、MDS の結果が以前の結果と大きく変わってしまう可能性があり、あまり適切な操作とは考えにくい。

よって、本研究では、収束計算の過程の各々を表示させることによって、ドラスティックな画面の変化が起きないようにする。また、MDS の結果は初期配置に依存するが、前回の MDS の配置を初期配置とすることによって、以前の MDS の結果と類似した結果を画面に表示させることが可能となった。

4. 応用事例：研究者グループの可視化

前節で提案した MDS のインタラクティブな操作を行う枠組みの有効性を検証するために、研究者グループの可視化を行うシステムを開発した。この可視化システムでは、次のような基準により研究者グループの可視化を行う。

- 研究者間の類似性は、研究者が著者として含まれる論文のアブストラクトの類似性によって判断する。ただし、アブストラクトでは、単語のバラエティが少なく、高頻度語を中心とした類似度が設定される可能性が高いため、確率型潜在意味解析 (PLSA)[5] を行うことで、あらかじめ次元の縮退を行った。
- 距離尺度学習に基づく MDS のインタラクティブな操作
  (1) ユーザは、同じグループに所属させたい研究者を複数選択し、この操作により、選択した著者全て同じラベルが与えられる。同じラベルを与えたい研究者がいる場合は、ラベルのついた研究者と一緒に選択し、ラベル付与を行うことにより、同一のラベルをつける。また、新たに複数の研究者を選択してラベルをつける場合には、以前と異なるラベルを与える。
  (2) ユーザは、収束計算のためのコントロールパラメータを調整することで、変化の度合いをコントロールする。
  (3) また、新しく得られた距離尺度の情報をユーザへのフィードバックのための情報として提供する。計算の結果得られたマハラノビス行列 M_t は、縮退した次元に関して有用な次元に関連するものは値が大きくなり、有用でない次元に関するものは値が小さくなることが想定される。
  この情報を、以下の式 18 を用いて、有効な情報を単語の情報に対応づける。具体的には、マハラノビス行列 M_t に対し、有用な次元については正、有用でない次元は負となるような重要度差分ベクトル $\Delta\vec{I_t}$ を作成し、PLSA における縮退次元と元の単語次元の関係を表す確率 $P(w|z_i)$(i 番目のトピックにおける w の出現確率) を組み合わせて、単語の重要度の変化 $\Delta\vec{I_{wt}}$ を計算する。これを各 PLSA 次元における関連度の高い単語 ($P(w|z_i)$ が高い単語) にについて計算したうえで、その値の上位のものを有用な単語のリストとして表示する。
  (4) ユーザは必要に応じ、(1) に戻る。この操作を繰り返す。

$$\vec{I_t} = M_t\vec{1} \quad (16)$$

$$\Delta\vec{I_t} = \vec{I_t} - \vec{1} \quad (17)$$

$$\Delta\vec{I_{wt}} = \sum_{i=0}^{T} \Delta\vec{I_t} P(w|z_i) \quad (18)$$

本研究を機械学習に関する論文アブストラクトデータベース (Pascal Visualization Challenge [⋆1]に適用することにより、研究者間の類似度の可視化を行った。具体的には、以下の手順により、研究者データを作成した。

(1) アブストラクト情報の作成
アブストラクトに対し、TreeTagger [⋆2] を用いて形態素解析し、名詞、形容詞、副詞、動詞を抽出した上で、ストップリスト (e.g., be, do, one, etc.) に含まれる語を削除し、アブストラクトに対するインデックスを作成した。

(2) 研究者情報の作成
全ての研究者について、その研究者を著者の一部に含むアブストラクトのインデックスを集約して、研究者情報のインデックスを作成した。インデックス語に対する重みとしては、TF · IDF を利用し、ベクトルの長さが 1 になるように正規化を行った。

(3) PLSA による次元圧縮による最終データ作成
最後に、得られたベクトルを PLSA を用いて次元を縮退させることにより、各研究者の特徴ベクトルを作成した。

⋆1 http://analytics.ijs.si/~blazf/pvc/data.html
⋆2 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Vol.2010-FI-98 No.8
Vol.2010-DD-75 No.8
2010/3/5

今回の実験で用いたデータには、313 人の研究者に対し、2986 語のインデックス語が存在した。これを 40 次元に縮退して分析を行った。

図 1 は MDS による可視化結果の一部を示す。ここで、ユーザは、“Michele Sebag”, “Mathieu Roche”, “Jerome Aze” の 3 名が同じグループに属するようにラベルを設定した (図 1 中の赤い点)。この 3 名は各々複数の論文がデータベースに登録されており、そのトピック毎の論文数は表 1 の通りである。一つの論文に複数のトピックが割り当てられていることもあるので、総計が以前のものの和と一致しない。

表 1 3 名の研究者の研究トピック毎の論文数 (重複あり)

名前	学習	自然言語処理	情報検索	アルゴリズム	総計
Michele Sebag	6	2	0	6	11
Jerome Aze	5	5	1	1	7
Mathieu Roche	2	7	1	0	7

この 3 名の研究者は、自然言語処理と学習の分野で 2 編の共著論文がある。この 3 名が類似しているという要求のもとで、距離尺度学習を行い、MDS で可視化した結果を図 2 に示す。

3 名の研究者の距離を小さくするように距離尺度学習が行われたため、3 名が近くに配置されることが確認された。この時の勾配に基づいて計算した分類に有用な単語の上位を表 2 に示す。また、実際に、この 3 名の研究者の周りにいる他の研究者をみると、“Francois Jacquenet” や “Marie-Jeanne Lesot” といったテキストマイニングに機械学習のアルゴリズムを応用するする研究者を近くに見つけることができるようになっており、ユーザの興味に応じた MDS の可視化結果の修正が行われたと考える。

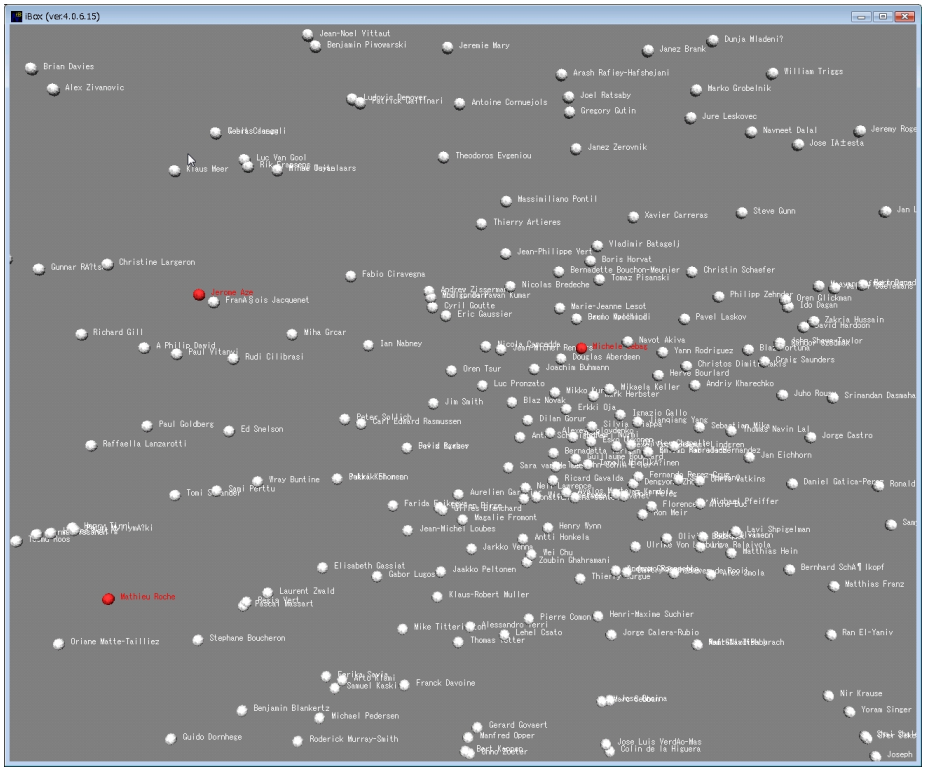

図 1 MDS による初期配置

5. Conclusions

本論文では、距離尺度学習の枠組みを利用することによって、多次元尺度構成法による可視化結果をインタラクティブに操作する方法を提案した。また、実際の研究者データの可視化に適用することによって、本システムが活用できることを確認した。

しかし、このようなシステムの有効性を判断するための基準を決めて、さらなる有効性に関する検討を行う必要がある。今後の展開としては、文書データ以外への展開についても検討したいと考えている。

Vol.2010-FI-98 No.8
Vol.2010-DD-75 No.8
2010/3/5

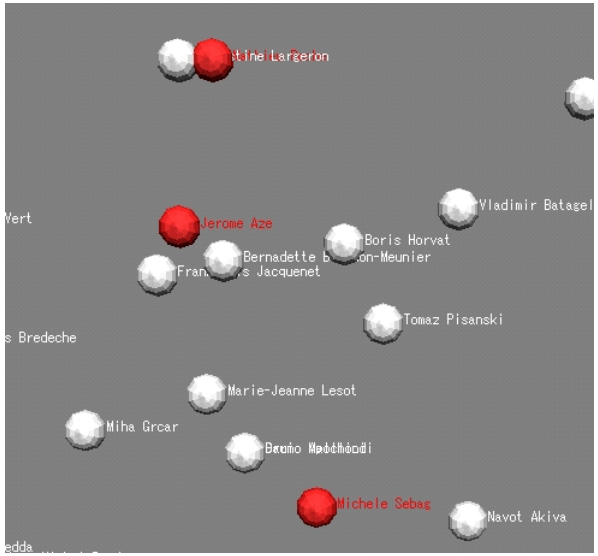

図 2 距離尺度学習による MDS 結果の修正

表 2 分類に有用なキーワード

キーワード	重要度 (10^{-11})
style	3.05
algorithm	2.44
problem	2.42
recognition	2.28
classification	2.11
number	1.85
different	1.67
musical	1.60
text	1.60
result	1.55

参 考 文 献

1) Buja, A., Swayne, D.F., Littman, M.L., Dean, N., Hofmann, H. and Chen, L.: Data Visualization With Multidimensional Scaling, *Journal of Computational and Graphical Statistics*, Vol.17, No.2, pp.444–472 (2008).
2) Broekens, J. and Cocx, T.: Object-Centered Interactive Multi-Dimensional Scaling: Ask the Expert, *Proceedings of the Eighteenth Belgium-Netherlands Conference on Artificial Intelligence (BNAIC2006)*, pp.59–66 (2006).
3) 相原健郎, 堀浩一, 大須賀節雄: 断片的な情報の集まりから知識を構築する過程の支援, 人工知能学会誌, Vol.11, No.3, pp.432–439 (1996).
4) Kilian Q.Weinberger, L. K.S.: Distance Metric Learning for Large Margin Nearest Neighbor Classification, *Journal of Machine Learning Research*, Vol.10, pp.207–244 (2009).
5) Hofmann, T.: Probabilistic latent semantic indexing, *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, ACM, pp.50–57 (1999).